

Estimating Conditional Quantiles with the Help of the Pinball Loss

Ingo Steinwart
Information Sciences Group CCS-3
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
ingo@lanl.gov

Andreas Christmann
University of Bayreuth
Department of Mathematics
D-95440 Bayreuth
Andreas.Christmann@uni-bayreuth.de

November 4, 2008

Abstract

Using the so-called pinball loss for estimating conditional quantiles is a well-known tool in both statistics and machine learning. So far, however, only little work has been done to quantify the efficiency of this tool for non-parametric (modified) empirical risk minimization approaches. The goal of this work is to fill this gap by establishing inequalities that describe how close approximate pinball risk minimizers are to the corresponding conditional quantile. These inequalities, which hold under mild assumptions on the data-generating distribution, are then used to establish so-called variance bounds, which recently turned out to play an important role in the statistical analysis of (modified) empirical risk minimization approaches. To illustrate the use of our new inequalities, we then utilize them to establish an oracle inequality for support vector machines that use the pinball loss. Here, it turns out that we obtain learning rates, which are optimal in a min-max sense under some standard assumptions on the regularity of the conditional quantile function.

1 Introduction

Let P be a distribution on $X \times Y$, where X is an arbitrary set equipped with a σ -algebra, and $Y \subset \mathbb{R}$ is closed. The goal of quantile regression is to estimate the conditional quantile, i.e., the set valued function

$$F_{\tau, P}^*(x) := \{t \in \mathbb{R} : P((-\infty, t] | x) \geq \tau \text{ and } P([t, \infty) | x) \geq 1 - \tau\}, \quad x \in X,$$

where $\tau \in (0, 1)$ is a fixed constant specifying the desired quantile level and $P(\cdot | x)$, $x \in X$, is the (regular) conditional probability of P . Let us assume for a moment¹ that $F_{\tau, P}^*(x)$ consists of singletons, i.e., there exists a function $f_{\tau, P}^* : X \rightarrow \mathbb{R}$, called the conditional τ -quantile function, such that $F_{\tau, P}^*(x) = \{f_{\tau, P}^*(x)\}$, $x \in X$. Then one approach to estimate the conditional τ -quantile function is based on the so-called τ -pinball loss $L : Y \times \mathbb{R} \rightarrow [0, \infty)$, which is defined by

$$L(y, t) := \begin{cases} (1 - \tau)(t - y) & \text{if } y < t \\ \tau(y - t) & \text{if } y \geq t. \end{cases}$$

With the help of this loss function we define the L -risk of a (measurable) function $f : X \rightarrow \mathbb{R}$ by

$$\mathcal{R}_{L, P}(f) := \mathbb{E}_{(x, y) \sim P} L(y, f(x)) = \int_{X \times Y} L(y, f(x)) dP(x, y).$$

Now recall that $f_{\tau, P}^*$ is up to P_X -zero sets the *only* function that minimizes the L -risk, i.e. $\mathcal{R}_{L, P}(f_{\tau, P}^*) = \inf \mathcal{R}_{L, P}(f) =: \mathcal{R}_{L, P}^*$, where the infimum is taken over all measurable functions $f : X \rightarrow \mathbb{R}$. Based on this observation several estimators minimizing a (modified) empirical L -risk were proposed (see [7] for a survey on both parametric and non-parametric methods) for situations where P is unknown, but i.i.d. samples $D := ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ drawn from P are given.

Empirical methods estimating quantile functions with the help of the pinball loss typically obtain functions f_D for which $\mathcal{R}_{L, P}(f_D)$ is close to $\mathcal{R}_{L, P}^*$ with high probability. In general, however, this only implies that f_D is close to $f_{\tau, P}^*$ in a very weak sense (see [13, Remark 3.18]), but recently, [15] established *self-calibration inequalities* of the form

$$\|f - f_{\tau, P}^*\|_{L_r(P_X)} \leq c_P \sqrt{\mathcal{R}_{L, P}(f) - \mathcal{R}_{L, P}^*}, \quad (1)$$

which holds under mild assumptions on P , which are described by the parameter $r \in (0, 1]$. The first goal of this paper is to generalize and improve these inequalities. Moreover, we will use these new self-calibration inequalities to establish *variance bounds* for the pinball risk, which in turn are known to improve the statistical analysis of empirical risk minimization approaches.

The second goal of this paper is to apply the self-calibration inequalities and the variance bounds to support vector machines (SVMs) for quantile regression. Recall, that [12, 6, 18] proposed an SVM that finds a solution $f_{D, \lambda} \in H$ of

$$\arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)), \quad (2)$$

where $\lambda > 0$ is a regularization parameter and H is a reproducing kernel Hilbert space (RKHS) over X . In [6, 18] it was worked out how to solve this optimization problem with numerical techniques, which are nowadays standard in the machine learning literature. Moreover, [18] also provided an exhaustive empirical study which shows the

¹Most of our main results later in this work do not require this assumption, but here in the introduction it makes the exposition more transparent.

excellent performance of this SVM approach. We have recently established an oracle inequality for these SVMs in [15], which were based on (1) and the resulting variance bounds. In this paper, we improve this oracle inequality with the help of the new self-calibration inequalities and variance bounds. It turns out that the resulting learning rates are substantially faster than those obtained in [15]. Finally, we briefly discuss an adaptive parameter selection strategy, which is based on a training/validation approach.

The rest of this paper is organized as follows: In Section 2, we present both our new self-calibration inequality and the new variance bound. We also introduce the assumptions on P that lead to these inequalities, and discuss how these inequalities improve our former results in [15]. In Section 3, we then use these new inequalities to establish an oracle inequality for the SVM approach above. In addition, we discuss the resulting learning rates and how these can be achieved in an adaptive way. Finally, all proofs are contained in Section 4.

2 Main results

In the following, X is an arbitrary, non-empty set equipped with a σ -algebra, and $Y \subset \mathbb{R}$ is a closed non-empty set. Given a distribution P on $X \times Y$ we further assume throughout this paper that the σ -algebra on X is complete with respect to the marginal distribution P_X of P , i.e., every subset of a P_X -zero set is contained in the σ -algebra. Since the latter can always be ensured by increasing the original σ -algebra in a suitable manner we note that this is not a restriction at all.

In order to formulate the main results of this section, we need to introduce some assumptions on the data-generating distribution P . To this end, let Q be a distribution on \mathbb{R} and $\tau \in (0, 1)$. Then the τ -quantile of Q is the set

$$F_\tau^*(Q) := \{t \in \mathbb{R} : Q((-\infty, t]) \geq \tau \text{ and } Q([t, \infty)) \geq 1 - \tau\}.$$

It is not hard to show that $F_\tau^*(Q)$ is a bounded and closed interval. We write

$$\begin{aligned} t_{\min}^*(Q) &:= \min F_\tau^*(Q) \\ t_{\max}^*(Q) &:= \max F_\tau^*(Q), \end{aligned}$$

and we usually omit the argument Q if the considered distribution is clearly determined from the context. We further need the following notion.

Definition 2.1 *A distribution Q on \mathbb{R} with support $\text{supp } Q \subset [-1, 1]$ is said to have a τ -quantile of type $q \in (1, \infty)$, if there exist constants $\alpha_Q \in (0, 2]$ and $b_Q > 0$ such that, for all $s \in [0, \alpha_Q]$, we have*

$$Q((t_{\min}^* - s, t_{\min}^*)) \geq b_Q s^{q-1} \tag{3}$$

$$Q((t_{\max}^*, t_{\max}^* + s)) \geq b_Q s^{q-1}. \tag{4}$$

Moreover, we say that Q has a τ -quantile of type $q = 1$, if both $Q(\{t_{\min}^*\}) > 0$ and $Q(\{t_{\max}^*\}) > 0$. In this case, we define $\alpha_Q := 2$ and

$$b_Q := \min\{Q(\{t_{\min}^*\}), Q(\{t_{\max}^*\})\}.$$

Finally, in both cases we define

$$\gamma_Q := b_Q \alpha_Q^{q-1}. \quad (5)$$

Leading examples for distributions having τ -quantiles of type $q = 2$ for $\tau \in (0, 1)$ are distributions Q with a Lebesgue density $h_Q(x) \geq b_Q > 0$ for all $x \in \text{supp } Q$. Moreover, distributions of type $q \neq 2$ can be realized by making appropriate assumptions on the behavior of h_Q around the quantile of interest. Finally, note that these distributions are *not* the only distributions of type q .

As outlined in the introduction, we are not interested in a single distribution Q on \mathbb{R} but in distributions P on $X \times \mathbb{R}$. The following definition extends the previous definition to such probability measures.

Definition 2.2 *Let $p \in (0, \infty]$ and $q \in [1, \infty)$. A distribution P on $X \times [-1, 1]$ is said to have a τ -quantile of p -average type q if for P_X -almost all $x \in X$ the conditional distribution $P(\cdot | x)$ has a τ -quantile of type q , and the function $\gamma : X \rightarrow [0, \infty]$ defined by*

$$\gamma(x) := \gamma_{P(\cdot | x)}, \quad x \in X,$$

where $\gamma_{P(\cdot | x)}$ is given by (5), satisfies $\gamma^{-1} \in L_p(P_X)$.

In the following theorem, which establishes the announced self-calibration inequality, we need the distance $\text{dist}(t, A)$ between an element $t \in \mathbb{R}$ and a subset $A \subset \mathbb{R}$, i.e., the quantity

$$\text{dist}(t, A) := \inf_{s \in A} |t - s|.$$

Moreover, $\text{dist}(f, F_{\tau, P}^*)$ denotes the function $x \mapsto \text{dist}(f(x), F_{\tau, P}^*(x))$. With these preparations the self-calibration inequality reads as follows.

Theorem 2.3 *Let L be the τ -pinball loss, $p \in (0, \infty]$ and $q \in [1, \infty)$ be real numbers, and $r := \frac{pq}{p+1}$. Moreover, let P be a distribution on $X \times [-1, 1]$ that has a τ -quantile of p -average type $q \in [1, \infty)$. Then for all $f : X \rightarrow [-1, 1]$ we have*

$$\|\text{dist}(f, F_{\tau, P}^*)\|_{L_r(P_X)} \leq 2^{1-1/q} q^{1/q} \|\gamma^{-1}\|_{L_p(P_X)}^{1/q} (\mathcal{R}_{L, P}(f) - \mathcal{R}_{L, P}^*)^{1/q}.$$

Let us briefly compare the self-calibration inequalities above with the ones we established in [15]. To this end, we can solely focus on the case $q = 2$, since this was the only case considered in [15]. For the same reason, we can restrict our considerations to distributions P that have a unique conditional τ -quantile $f_{\tau, P}^*(x)$ for P_X -almost all $x \in X$. Then Theorem 2.3 yields

$$\|f - f_{\tau, P}^*\|_{L_r(P_X)} \leq 2 \|\gamma^{-1}\|_{L_p(P_X)}^{1/2} (\mathcal{R}_{L, P}(f) - \mathcal{R}_{L, P}^*)^{1/2}$$

for $r := \frac{2p}{p+1}$. On the other hand, it was shown in [15] that

$$\|f - f_{\tau, P}^*\|_{L_{r/2}(P_X)} \leq \sqrt{2} \|\gamma^{-1}\|_{L_p(P_X)}^{1/2} (\mathcal{R}_{L, P}(f) - \mathcal{R}_{L, P}^*)^{1/2}$$

under the *additional* assumption that the conditional widths $\alpha_{P(\cdot|x)}$ considered in Definition 2.1 are *independent* from x . This shows that our new self-calibration inequality is more general and, modulo the constant $\sqrt{2}$, also stronger.

It is well-known that self-calibration inequalities for Lipschitz continuous losses lead to variance bounds, which in turn are important for the statistical analysis of empirical risk minimization approaches, see [8, 9, 10, 11, 1, 2]. For the pinball loss, the self-calibration inequality established above leads to the following variance bound.

Theorem 2.4 *Let L be the τ -pinball loss, $p \in (0, \infty]$ and $q \in [1, \infty)$ be real numbers, and*

$$\vartheta := \min\left\{\frac{2}{q}, \frac{p}{p+1}\right\}.$$

Moreover, let P be a distribution on $X \times [-1, 1]$ that has a τ -quantile of p -average type q . Then for all $f : X \rightarrow [-1, 1]$ there exists a function $f_{\tau, P}^ : X \rightarrow [-1, 1]$ with $f_{\tau, P}^*(x) \in F_{\tau, P}^*(x)$ for P_X -almost all $x \in X$ such that*

$$\mathbb{E}_P(L \circ f - L \circ f_{\tau, P}^*)^2 \leq 2^{2-\vartheta} q^\vartheta \|\gamma^{-1}\|_{L_p(P_X)}^\vartheta (\mathcal{R}_{L, P}(f) - \mathcal{R}_{L, P}^*)^\vartheta,$$

where we used the shorthand $L \circ f$ for the function $(x, y) \mapsto L(y, f(x))$.

Again, it is straightforward to show that the variance bound above is both more general and stronger than the variance bound we established in [15]. We omit the details for the sake of brevity.

3 An Application to Support Vector Machines

The goal of this section is to establish an oracle inequality for the SVM defined in (2). The use of this oracle inequality is then illustrated by some learning rates we derive from it.

Let us begin by recalling some RKHS theory (see, e.g., [16, Chapter 4] for a more detailed account). To this end, let $k : X \times X \rightarrow \mathbb{R}$ be a measurable kernel, i.e., a measurable function that is symmetric and positive definite. Then the associated RKHS H consists of measurable functions. Let us additionally assume that k is bounded with $\|k\|_\infty := \sup_{x \in X} \sqrt{k(x, x)} \leq 1$, which in turn implies that H consists of bounded functions and $\|f\|_\infty \leq \|f\|_H$ for all $f \in H$.

Suppose now that we have a distribution P on $X \times Y$. To describe the approximation error of SVMs we need the *approximation error function*

$$A(\lambda) := \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L, P}(f) - \mathcal{R}_{L, P}^*, \quad \lambda > 0,$$

where L is the τ -pinball loss. Recall that [16, Lemma 5.15 and Theorem 5.31], see also [17], showed $\lim_{\lambda \rightarrow 0} A(\lambda) = 0$, if the RKHS H is dense in $L_1(P_X)$, and the speed of this converge describes how well H approximates the Bayes L -risk. In particular, [16, Corollary 5.18] shows that $A(\lambda) \leq c\lambda$ for some constant $c > 0$ and all $\lambda > 0$ if and only if there exists an $f \in H$ such that $f(x) \in F_{\tau, P}^*(x)$ for P_X -almost all $x \in X$.

In order to describe the capacity of the RKHS H we further need the integral operator $T_k : L_2(P_X) \rightarrow L_2(P_X)$ that is defined by

$$T_k f(\cdot) := \int_X k(x, \cdot) f(x) dP_X(x), \quad f \in L_2(P_X).$$

It is well-known that this integral operator is self-adjoint and nuclear, see, e.g., [16, Theorem 4.27]. Consequently, it has at most countably many eigenvalues (including geometric multiplicities), which are all non-negative, and which, as a sequence, are summable. In the following we order these eigenvalues $\lambda_i(T_k)$. Moreover, if we only have finitely many eigenvalues we extend this finite sequence by zeros. As a result, we always can deal with a decreasing, non-negative sequence $\lambda_1(T_k) \geq \lambda_2(T_k) \geq \dots$ which satisfies

$$\sum_{i=1}^{\infty} \lambda_i(T_k) < \infty.$$

The finiteness of this sum can already be used to establish oracle inequalities, see e.g. [16, Theorem 7.22], but in the following we assume that the eigenvalues converge even faster to zero, since *a*) this case is satisfied for many RKHSs and *b*) it leads to better oracle inequalities. To be more precise, we assume that there exist constants $a \geq 1$ and $\varrho \in (0, 1)$ such that

$$\lambda_i(T_k) \leq a i^{-1/\varrho}, \quad i \geq 1. \quad (6)$$

One can show that this eigenvalue assumption is equivalent to an entropy number assumption on the inclusion $\text{id} : H \rightarrow L_2(P_X)$. Namely, (6) is satisfied if and only if we have

$$e_i(\text{id} : H \rightarrow L_2(P_X)) \leq \sqrt{a} i^{-1/(2\varrho)}, \quad i \geq 1,$$

where $e_i(S)$ denotes the i -th (dyadic) entropy numbers of a bounded linear operator S . We refer to [4] for information regarding entropy numbers and to [14] for a brief argument for this equivalence.

Finally, we also need the clipping operation, which, for fixed $M > 0$, is defined by

$$\hat{t} := \begin{cases} -M & \text{if } t < -M \\ t & \text{if } t \in [-M, M] \\ M & \text{if } t > M \end{cases} \quad (7)$$

for all $t \in \mathbb{R}$. In the following, we assume that $M := 1$ if not stated otherwise. With the help of these notations we can now formulate the following oracle inequality which is just a particular case of a more general inequality established in [16, Theorem 7.23].

Theorem 3.1 *Let L be the τ -pinball loss and P be a distribution on $X \times [-1, 1]$ for which there exists a function $f_{\tau, P}^* : X \rightarrow \mathbb{R}$ with $f_{\tau, P}^*(x) \in F_{\tau, P}^*(x)$ for P_X -almost all $x \in X$. We further assume that there exist constants $V \geq 2^{2-\vartheta}$ and $\vartheta \in [0, 1]$ such that*

$$\mathbb{E}_P(L \circ f - L \circ f_{\tau, P}^*)^2 \leq V (\mathcal{R}_{L, P}(f) - \mathcal{R}_{L, P}^*)^{\vartheta} \quad (8)$$

for all $f : X \rightarrow [-1, 1]$. Moreover, let H be a RKHS over X with bounded measurable kernel satisfying $\|k\|_\infty \leq 1$. In addition, assume that (6) is satisfied. Then there exists a constant K depending only on ϱ , V , and ϱ such that for all $\varsigma \geq 1$, $n \geq 1$, and $\lambda > 0$ we have with probability not less than $1 - 3e^{-\varsigma}$ that

$$\begin{aligned} \mathcal{R}_{L,P}(\widehat{f}_{D,\lambda}) - \mathcal{R}_{L,P}^* &\leq 9A(\lambda) + 30\sqrt{\frac{A(\lambda)}{\lambda}} \frac{\varsigma}{n} + K\left(\frac{a^\varrho}{\lambda^\varrho n}\right)^{\frac{1}{2-\varrho-\vartheta+\vartheta\varrho}} \\ &\quad + 3\left(\frac{72V\varsigma}{n}\right)^{\frac{1}{2-\vartheta}}. \end{aligned}$$

Let us now illustrate how this oracle inequality can be used to establish learning rates for estimating conditional quantiles. To this end, we assume that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that $A(\lambda) \leq c\lambda^\beta$ for all $\lambda > 0$. Then it is easy to show, see [16, Lemma A.1.7], that $\mathcal{R}_{L,P}(\widehat{f}_{D,\lambda_n})$ converges to $\mathcal{R}_{L,P}^*$ with rate $n^{-\gamma}$, where

$$\gamma := \min\left\{\frac{\beta}{\beta(2-\vartheta+\varrho\vartheta-\varrho)+\varrho}, \frac{2\beta}{\beta+1}\right\}, \quad (9)$$

provided that we have chosen λ by $\lambda_n = n^{-\gamma/\beta}$. Note that this choice of λ yields the best learning rates from Theorem 3.1. Unfortunately, however, this choice requires to know the usually unknown parameters β , ϑ , and ϱ . On the other hand, [16, Theorem 7.24] shows that this rate can also be achieved by selecting λ in a data-dependent way with the help of a validation data set. In other words, the learning rates above can be achieved without knowing the existence of the above parameters nor their particular values.

Let us now consider how these learning rates in terms of risks translate into rates for

$$\|\widehat{f}_{D,\lambda_n} - f_{\tau,P}^*\|_{L_r(P_X)}.$$

To this end, we assume that P has a τ -quantile of p -average type q , where we additionally assume for the sake of simplicity that $r := \frac{pq}{p+1} \leq 2$. Note that the latter is satisfied for all p if $q \leq 2$, i.e., if all conditional distributions are concentrated around the quantile as least as much as the uniform distribution. We refer to the discussion following Definition 2.1 for a precise statement. Moreover, we additionally assume that the conditional quantiles $F_{\tau,P}^*(x)$ are singletons for P_X -almost all $x \in X$. Then Theorem 2.4 provides a variance bound of the form (8) for $\vartheta := p/(p+1)$, and hence γ defined in (9) becomes

$$\gamma = \min\left\{\frac{\beta(p+1)}{\beta(2+p-\varrho)+\varrho(p+1)}, \frac{2\beta}{\beta+1}\right\}$$

By Theorem 2.3 we consequently see that $\|\widehat{f}_{D,\lambda_n} - f_{\tau,P}^*\|_{L_r(P_X)}$ converges with rate $n^{-\gamma/q}$ to zero, where $r := pq/(p+1)$. To illustrate this learning rate, let us assume that we have picked a RKHS H with $f_{\tau,P}^* \in H$. Then we have $\beta = 1$, and hence it is easy to check that the latter learning rate reduces to

$$n^{-\frac{p+1}{q(2+p+ep)}}.$$

For the sake of simplicity, let us further assume that the conditional distributions do not change too much in the sense that $p = \infty$. Then we have $r = q$ and the learning rate for $\|\widehat{f}_{D, \lambda_n} - f_{\tau, P}^*\|_{L_q(P_X)}$ becomes

$$n^{-\frac{1}{q(1+\varrho)}}.$$

In other words,

$$\int_X |\widehat{f}_{D, \lambda_n} - f_{\tau, P}^*|^q dP_X \quad (10)$$

converges with rate $n^{-1/(1+\varrho)}$. The latter shows that the value of q does not change the learning rate for (10), but only the exponent in (10). Now note that by our assumption on P and the definition of the clipping operation we have

$$\|\widehat{f}_{D, \lambda_n} - f_{\tau, P}^*\|_\infty \leq 2,$$

and consequently small values of q emphasize the discrepancy of $\widehat{f}_{D, \lambda_n}$ to $f_{\tau, P}^*$ more than large values of q do. In this sense, a stronger average concentration around the quantile of interest is helpful for the learning process.

Let us now have a closer look to the special case $q = 2$, which is probably the most interesting case for applications. Then we have the learning rate $n^{-1/(2(1+\varrho))}$ for

$$\|\widehat{f}_{D, \lambda_n} - f_{\tau, P}^*\|_{L_2(P_X)}.$$

Now recall that the conditional median equals the conditional mean for *symmetric* conditional distributions $P(\cdot|x)$. Moreover, if H is a Sobolev space $W^m(X)$, where $m > d/2$ denotes the smoothness index and X is the unit ball in \mathbb{R}^d , then H consists of continuous functions, and [5] shows that H satisfies (6) for $\varrho := d/(2m)$ and P_X being the uniform distribution on X . Consequently, we see that in this case the latter convergence rate is optimal in a min-max sense. Finally, recall that in the case $\beta = 1$, $q = 2$, and $p = \infty$ discussed so far, the results derived by [15] only yielded a learning rate of $n^{-1/(3(1+\varrho))}$ for

$$\|\widehat{f}_{D, \lambda_n} - f_{\tau, P}^*\|_{L_1(P_X)}.$$

In other words, the earlier rates from [15] are not only worse by a factor of $3/2$ in the exponent but also stated in terms of the weaker $L_1(P_X)$ -norm. In addition, [15] only considered the case $q = 2$, and hence we see that our new results are also more general.

4 Proofs

Let us first recall some notions from [13] and [16, Chapter 3] which investigated surrogate losses in general and the question of how approximate risk minimizers approximate exact risk minimizers in particular. To this end, let $L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)$ be

a measurable function, which we call a loss in the following. For a distribution P and an $f : X \rightarrow \mathbb{R}$ the L -risk is then defined by

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L(x, y, f(x)) dP(x, y),$$

and, as usual, the Bayes L -risk, is denoted by $\mathcal{R}_{L,P}^* := \inf \mathcal{R}_{L,P}(f)$, where the infimum is taken over all measurable functions $f : X \rightarrow \mathbb{R}$. In addition, given a distribution Q on Y the *inner L -risks* were defined by

$$\mathcal{C}_{L,Q,x}(t) := \int_Y L(x, y, t) dQ(y), \quad x \in X, t \in \mathbb{R},$$

and the *minimal inner L -risks* were denoted by $\mathcal{C}_{L,Q,x}^* := \inf \mathcal{C}_{L,Q,x}(t)$, $x \in X$, where the infimum is taken over all $t \in \mathbb{R}$. Moreover, following [13] we usually omit the indexes x or Q if L is independent of x or y , respectively. Obviously, we have

$$\mathcal{R}_{L,P}(f) = \int_X \mathcal{C}_{L,P(\cdot|x),x}(f(x)) dP_X(x), \quad (11)$$

and [13, Theorem 3.2] further shows that $x \mapsto \mathcal{C}_{L,P(\cdot|x),x}^*$ is measurable if the σ -algebra on X is complete. In this case, it was also shown that the intuitive formula

$$\mathcal{R}_{L,P}^* = \int_X \mathcal{C}_{L,P(\cdot|x),x}^* dP_X(x)$$

holds, i.e. the Bayes L -risk is obtained by minimizing the inner risks and subsequently integrating with respect to the marginal distribution P_X . Based on this observation the basic idea in [13] is to consider both steps separately. In particular, it turned out that the sets of ε -approximate minimizers

$$\mathcal{M}_{L,Q,x}(\varepsilon) := \{t \in \mathbb{R} : \mathcal{C}_{L,Q,x}(t) < \mathcal{C}_{L,Q,x}^* + \varepsilon\}, \quad \varepsilon \in [0, \infty],$$

and the set of *exact minimizers*

$$\mathcal{M}_{L,Q,x}(0^+) := \bigcap_{\varepsilon > 0} \mathcal{M}_{L,Q,x}(\varepsilon)$$

play a crucial role. As in [13] we again omit the subscripts x and Q in these definitions if L happens to be independent of x or y , respectively.

Let us now compute the excess inner risks and the set of exact minimizers for the pinball loss. To this end recall (see, e.g., [3, Theorem 23.8]) that given a distribution Q on \mathbb{R} and a *non-negative* measurable function $g : X \rightarrow [0, \infty)$ we have

$$\int_{\mathbb{R}} g dQ = \int_0^\infty Q(g \geq s) ds. \quad (12)$$

With the help of these preparations we can now show the following preliminary result, which is a generalization of [16, Proposition 3.9].

Proposition 4.1 *Let L be the τ -pinball loss and Q be a distribution on \mathbb{R} with $\mathcal{C}_{L,Q}^* < \infty$. Then there exist $q_+, q_- \in [0, 1]$ with $q_+ + q_- = Q([t_{\min}^*, t_{\max}^*])$, and for all $t \geq 0$ we have*

$$\mathcal{C}_{L,Q}(t_{\max}^* + t) - \mathcal{C}_{L,Q}^* = tq_+ + \int_0^t Q((t_{\max}^*, t_{\max}^* + s)) ds, \quad (13)$$

$$\mathcal{C}_{L,Q}(t_{\min}^* - t) - \mathcal{C}_{L,Q}^* = tq_- + \int_0^t Q((t_{\min}^* - s, t_{\min}^*)) ds. \quad (14)$$

In addition, we have $\mathcal{M}_{L,Q}(0^+) = F_\tau^*(Q)$.

Proof: Obviously, we have $Q((-\infty, t_{\max}^*]) + Q([t_{\max}^*, \infty)) = 1 + Q(\{t_{\max}^*\})$, and hence we obtain $\tau \leq Q((-\infty, t_{\max}^*]) \leq \tau + Q(\{t_{\max}^*\})$. In other words, there exists a $q_+ \in [0, 1]$ satisfying $0 \leq q_+ \leq Q(\{t_{\max}^*\})$ and

$$Q((-\infty, t_{\max}^*]) = \tau + q_+. \quad (15)$$

Let us consider the distribution \tilde{Q} defined by $\tilde{Q}(A) := Q(t_{\max}^* + A)$ for all measurable $A \subset \mathbb{R}$. Then it is not hard to see that $t_{\max}^*(\tilde{Q}) = 0$. Moreover, we obviously have $\mathcal{C}_{L,Q}(t_{\max}^* + t) = \mathcal{C}_{L,\tilde{Q}}(t)$ for all $t \in \mathbb{R}$. Let us now compute the inner risks of L with respect to \tilde{Q} . To this end, we fix a $t \geq 0$. Then we have

$$\int_{y < t} (y - t) d\tilde{Q}(y) = \int_{y < 0} y d\tilde{Q}(y) - t\tilde{Q}((-\infty, t)) + \int_{0 \leq y < t} y d\tilde{Q}(y)$$

and

$$\int_{y \geq t} (y - t) d\tilde{Q}(y) = \int_{y \geq 0} y d\tilde{Q}(y) - t\tilde{Q}([t, \infty)) - \int_{0 \leq y < t} y d\tilde{Q}(y)$$

and hence we obtain

$$\begin{aligned} \mathcal{C}_{\tilde{Q},L}(t) &= (\tau - 1) \int_{y < t} (y - t) d\tilde{Q}(y) + \tau \int_{y \geq t} (y - t) d\tilde{Q}(y) \\ &= \mathcal{C}_{\tilde{Q},L}(0) - \tau t + t\tilde{Q}((-\infty, 0)) + t\tilde{Q}([0, t)) - \int_{0 \leq y < t} y d\tilde{Q}(y). \end{aligned}$$

Moreover, using (12) we find

$$\begin{aligned} t\tilde{Q}([0, t)) - \int_{0 \leq y < t} y d\tilde{Q}(y) &= \int_0^t \tilde{Q}([0, t)) ds - \int_0^t \tilde{Q}([s, t)) ds \\ &= t\tilde{Q}(\{0\}) + \int_0^t \tilde{Q}((0, s)) ds, \end{aligned}$$

and since (15) implies $\tilde{Q}((-\infty, 0)) + \tilde{Q}(\{0\}) = \tilde{Q}((-\infty, 0]) = \tau + q_+$ we thus obtain (13). Now (14) can be derived from (13) by considering the pinball loss with parameter $1 - \tau$ and the distribution \bar{Q} defined by $\bar{Q}(A) := Q(-t_{\min}^* - A)$, $A \subset \mathbb{R}$ measurable. This further yields a q_- satisfying $0 \leq q_- \leq Q(\{t_{\min}^*\})$ and $Q([t_{\min}^*, \infty)) = 1 - \tau + q_-$. By (15) we then find $q_+ + q_- = Q([t_{\min}^*, t_{\max}^*])$. The final assertion now easily follows from the formulas for the excess risks, and is, in addition, also well-known. \blacksquare

For the proof of Theorem 2.3 we need to recall a few more concepts from [13] or [16, Chapter 3]. To this end, let us now assume that our loss is independent of x , i.e. we consider a measurable function $L : Y \times \mathbb{R} \rightarrow [0, \infty]$. We write

$$\mathcal{Q}_{\min}(L) := \{Q : Q \text{ is a distribution on } \mathbb{R} \text{ such that } \mathcal{M}_{L,Q}(0^+) \neq \emptyset\},$$

i.e. $\mathcal{Q}_{\min}(L)$ contains the distributions on \mathbb{R} whose inner L -risks have at least one exact minimizer. Furthermore, note that this definition immediately yields $\mathcal{C}_{L,Q}^* < \infty$ for all $Q \in \mathcal{Q}_{\min}(L)$. Following [13] we now define the *self-calibration loss* of L by

$$\check{L}(Q, t) := \inf_{t^* \in \mathcal{M}_{L,Q}(0^+)} |t - t^*|, \quad Q \in \mathcal{Q}_{\min}(L), t \in \mathbb{R}. \quad (16)$$

This loss is a *template* loss in the sense of [13] or [16, Chapter 3], i.e., for a given distribution P on $X \times Y$, where X has a complete σ -algebra and $P(\cdot | x) \in \mathcal{Q}_{\min}(L)$ for P_X -almost all $x \in X$, the P -instance

$$\check{L}_P(x, t) := \check{L}(P(\cdot | x), t), \quad x \in X, t \in \mathbb{R},$$

is measurable, and hence a loss. [13] or [16, Chapter 3] extended the definition of inner risks to the self-calibration loss by setting $\mathcal{C}_{\check{L},Q}(t) := \check{L}(Q, t)$, and based on this, the minimal inner risks and their (approximate) minimizers were defined in the obvious ways. Moreover, the *self-calibration function* was defined by

$$\delta_{\max, \check{L}, L}(\varepsilon, Q) := \text{dist}(t, \mathcal{M}_{L,Q}(0^+)) := \inf_{t \in \mathbb{R} : \check{L}(Q, t) \geq \varepsilon} \mathcal{C}_{L,Q}(t) - \mathcal{C}_{L,Q}^*.$$

As shown in [13] or [16, Chapter 3.9] the self-calibration function satisfies

$$\delta_{\max, \check{L}, L}(\text{dist}(t, \mathcal{M}_{L,Q}(0^+)), Q) \leq \mathcal{C}_{L,Q}(t) - \mathcal{C}_{L,Q}^*, \quad t \in \mathbb{R}, \quad (17)$$

i.e. it measures how well an ε -approximate L -risk minimizer t approximate the set of exact L -risk minimizers.

Our next goal is to estimate the self-calibration function for the pinball loss. To this end we need the following simple technical lemma.

Lemma 4.2 *For $\alpha \in [0, 2]$ and $q \in [1, \infty)$ consider the function $\delta : [0, 2] \rightarrow [0, \infty)$ defined by*

$$\delta(\varepsilon) := \begin{cases} \varepsilon^q & \text{if } \varepsilon \in [0, \alpha] \\ q\alpha^{q-1}\varepsilon - \alpha^q(q-1) & \text{if } \varepsilon \in [\alpha, 2]. \end{cases}$$

Then for all $\varepsilon \in [0, 2]$ we have

$$\delta(\varepsilon) \geq \left(\frac{\alpha}{2}\right)^{q-1} \varepsilon^q.$$

Proof: Since $\alpha \leq 2$ and $q \geq 1$ we easily see by the definition of δ that the assertion is true for $\varepsilon \in [0, \alpha]$. Now consider the function $h : [0, 2] \rightarrow \mathbb{R}$ defined by

$$h(\varepsilon) := q\alpha^{q-1}\varepsilon - \alpha^q(q-1) - \left(\frac{\alpha}{2}\right)^{q-1} \varepsilon^q$$

for all $\varepsilon \in [0, 2]$. Obviously, it suffices to show that $h(\varepsilon) \geq 0$ for all $\varepsilon \in [\alpha, 2]$. To show the latter we first check that

$$h'(\varepsilon) = q\alpha^{q-1} - q\left(\frac{\alpha}{2}\right)^{q-1}\varepsilon^{q-1},$$

and hence we have $h'(\varepsilon) \geq 0$ for all $\varepsilon \in [0, 2]$. From this, $\alpha \leq 2$, and

$$h(\alpha) = \alpha^q - \left(\frac{\alpha}{2}\right)^{q-1}\alpha^q = \alpha^q\left(1 - \left(\frac{\alpha}{2}\right)^{q-1}\right) \geq 0$$

we then obtain the assertion. \blacksquare

Lemma 4.3 *Let L be the τ -pinball loss and Q be a distribution on \mathbb{R} with $\text{supp } Q \subset [-1, 1]$ that has a τ -quantile of type $q \in [1, \infty)$. Moreover, let $\alpha_Q \in (0, 2]$ and $b_Q > 0$ denote the corresponding constants. Then for all $\varepsilon \in [0, 2]$ we have*

$$\delta_{\max, \check{L}, L}(\varepsilon, Q) \geq q^{-1}b_Q\left(\frac{\alpha_Q}{2}\right)^{q-1}\varepsilon^q.$$

Proof: Obviously, the map $t \mapsto \mathcal{C}_{L, Q}(t) - \mathcal{C}_{L, Q}^*$ is convex, and thus it is decreasing on $(-\infty, t_{\min}^*]$ and increasing on $[t_{\max}^*, \infty)$. Since $\mathcal{M}_{L, Q}(0^+) = F_{\tau}^*(Q)$ is an interval, we hence find

$$\mathcal{M}_{\check{L}, Q}(\varepsilon) = \{t \in \mathbb{R} : \check{L}(Q, t) < \varepsilon\} = (t_{\min}^* - \varepsilon, t_{\max}^* + \varepsilon)$$

for all $\varepsilon > 0$. This gives

$$\begin{aligned} \delta_{\max, \check{L}, L}(\varepsilon, Q) &= \inf_{t \notin \mathcal{M}_{\check{L}, Q}(\varepsilon)} \mathcal{C}_{L, Q}(t) - \mathcal{C}_{L, Q}^* \\ &= \min\left\{\mathcal{C}_{L, Q}(t_{\min}^* - \varepsilon), \mathcal{C}_{L, Q}(t_{\max}^* + \varepsilon)\right\} - \mathcal{C}_{L, Q}^*. \end{aligned} \quad (18)$$

Let us first consider the case $q \in (1, \infty)$. For $t \in [0, \alpha_Q]$, Equations (13) and (4) then yield

$$\begin{aligned} \mathcal{C}_{L, Q}(t_{\max}^* + t) - \mathcal{C}_{L, Q}^* &= tq_+ + \int_0^t Q((t_{\max}^*, t_{\max}^* + s)) ds \\ &\geq b_Q \int_0^t s^{q-1} ds \\ &= q^{-1}b_Q t^q. \end{aligned}$$

In addition, for $q = 1$ this inequality follows in a similar fashion from (13), and an analogue estimate for $\mathcal{C}_{L, Q}(t_{\min}^* - t) - \mathcal{C}_{L, Q}^*$ can be shown by using (14) and (3). Having established these inequalities we then conclude by (18) that

$$\delta_{\max, \check{L}, L}(\varepsilon, Q) \geq q^{-1}b_Q\varepsilon^q$$

for all $\varepsilon \in [0, \alpha_Q]$. Now the assertion follows from Lemma 4.2. \blacksquare

In the following, we say that a distribution P on $X \times Y$ is of type \mathcal{Q} , where \mathcal{Q} is some set of distributions on \mathbb{R} , if $P(\cdot|x) \in \mathcal{Q}$ for P_X -almost all $x \in X$. Now we can formulate our last auxiliary result, which establishes a general self-calibration inequality.

Proposition 4.4 *Let $M > 0$ and $L : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty)$ be a convex loss such that for every $y \in [-M, M]$ the function $L(y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ has at least one global minimizer that is contained in $[-M, M]$. Moreover, let P be a type $\mathcal{Q}_{\min}(L)$ distribution on $X \times [-M, M]$ with $\mathcal{R}_{L,P}^* < \infty$. Assume that there exist $p \in (0, \infty]$, $q > 0$, and a function $\gamma : X \rightarrow [0, \infty]$ with $\gamma^{-1} \in \mathcal{L}_p(P_X)$ and*

$$\delta_{\max, \check{L}_P, L}(\varepsilon, P(\cdot|x), x) \geq \gamma(x) \varepsilon^q, \quad \varepsilon \in [0, 2M], x \in X.$$

Then for all measurable $f : X \rightarrow [-M, M]$ we have

$$\left(\int_X (\check{L}_P(x, f(x)))^{\frac{pq}{p+1}} dP_X(x) \right)^{\frac{p+1}{pq}} \leq \|\gamma^{-1}\|_{\mathcal{L}_p(P_X)}^{\frac{1}{q}} \left(\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^* \right)^{\frac{1}{q}}.$$

Proof: For $y \in [-M, M]$, we denote the set of minimizers of $L(y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ by $\mathcal{M}_y := \{t^* \in \mathbb{R} : L(y, t^*) = \inf_{t \in \mathbb{R}} L(y, t)\}$. Note that the convexity of L implies that \mathcal{M}_y is a closed interval. Moreover, by our assumptions we have $\mathcal{M}_y \cap [-M, M] \neq \emptyset$, and hence we have $\inf \mathcal{M}_y \leq M$ and $\sup \mathcal{M}_y \geq -M$. In addition, the convexity of L shows that $L(y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ is increasing on $[\sup \mathcal{M}_y, \infty)$ and decreasing on $(-\infty, \inf \mathcal{M}_y]$. Hence we have

$$L(y, \hat{t}) \leq L(y, t), \quad y \in [-M, M], t \in \mathbb{R},$$

where \hat{t} denotes the clipped value of t at $\pm M$, see (7). From this it is easy to conclude that

$$\mathcal{C}_{L,Q}(\hat{t}) \leq \mathcal{C}_{L,Q}(t)$$

for all $t \in \mathbb{R}$ and all distributions Q whose support is contained in $[-M, M]$. Consequently, we have $\mathcal{M}_{L,Q}(0^+) \cap [-M, M] \neq \emptyset$ for all such Q , which in turn implies that

$$\check{L}(Q, t) \leq 2M, \quad t \in [-M, M].$$

Now the assertion follows from (17) and the proof of [13, Proposition 3.19]. \blacksquare

Proof of Theorem 2.3: By Lemma 4.3 we obtain a lower bound on the self-calibration function of the pinball loss. Moreover, P is a type $\mathcal{Q}_{\min}(L)$ distribution, since the conditional minimizers, which are the conditional τ -quantiles, do exist. Now the assertion follows from Proposition 4.4. \blacksquare

Proof of Theorem 2.4: Let $f : X \rightarrow [-1, 1]$ be a measurable function and $f_{\tau,P}^* : X \rightarrow [-1, 1]$ be the P_X -almost surely uniquely determined measurable function that satisfies both

$$\begin{aligned} f_{\tau,P}^*(x) &\in F_{\tau,P}^*(x) \\ |f(x) - f_{\tau,P}^*(x)| &= \text{dist}(f(x), F_{\tau,P}^*(x)) \end{aligned}$$

for P_X -almost all $x \in X$. Let us write $r := \frac{pq}{p+1}$. We first consider the case $r \leq 2$, i.e., $\frac{2}{q} \leq \frac{p}{p+1}$. Using the Lipschitz continuity of L and Theorem 2.3 we then obtain

$$\begin{aligned} \mathbb{E}_P(L \circ f - L \circ f_{\tau,P}^*)^2 &\leq \mathbb{E}_{P_X}|f - f_{\tau,P}^*|^2 \\ &\leq \|f - f_{\tau,P}^*\|_\infty^{2-r} \mathbb{E}_{P_X}|f - f_{\tau,P}^*|^r \\ &\leq 2^{2-r/q} q^{r/q} \|\gamma^{-1}\|_{L_p(P_X)}^{r/q} (\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*)^{r/q}. \end{aligned}$$

Since $\frac{r}{q} = \frac{p}{p+1} = \vartheta$, we thus obtain the assertion in this case. Let us now consider the case $r > 2$. The Lipschitz continuity of L and Theorem 2.3 then yields

$$\begin{aligned} \mathbb{E}_P(L \circ f - L \circ f_{\tau,P}^*)^2 &\leq \left(\mathbb{E}_P(L \circ f - L \circ f_{\tau,P}^*)^r \right)^{2/r} \\ &\leq \left(\mathbb{E}_{P_X}|f - f_{\tau,P}^*|^r \right)^{2/r} \\ &\leq \left(2^{1-1/q} q^{1/q} \|\gamma^{-1}\|_{L_p(P_X)}^{1/q} (\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*)^{1/q} \right)^2 \\ &= 2^{2-2/q} q^{2/q} \|\gamma^{-1}\|_{L_p(P_X)}^{2/q} (\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*)^{2/q}. \end{aligned}$$

Since for $r > 2$ we have $\vartheta = 2/q$ we again obtain the assertion. \blacksquare

References

- [1] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33:1497–1537, 2005.
- [2] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101:138–156, 2006.
- [3] H. Bauer. *Measure and Integration Theory*. De Gruyter, Berlin, 2001.
- [4] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.
- [5] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.
- [6] C. Hwang and J. Shim. A simple quantile regression via support vector machine. In *Advances in Natural Computation: First International Conference (ICNC)*, pages 512–520. Springer, 2005.
- [7] R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.
- [8] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999.
- [9] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse, VI. Sr., Math.*, 9:245–303, 2000.

- [10] S. Mendelson. Geometric methods in the analysis of Glivenko-Cantelli classes. In D. Helmbold and B. Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 256–272. Springer, New York, 2001.
- [11] S. Mendelson. Learning relatively small classes. In D. Helmbold and B. Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 273–288. Springer, New York, 2001.
- [12] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Comput.*, 12:1207–1245, 2000.
- [13] I. Steinwart. How to compare different loss functions. *Constr. Approx.*, 26:225–287, 2007.
- [14] I. Steinwart. Some bounds on random entropy numbers with an application to support vector machines. Technical report, Los Alamos National Laboratory, 2008.
- [15] I. Steinwart and A. Christmann. How SVMs can estimate quantiles and the median. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 305–312. MIT Press, Cambridge, MA, 2008.
- [16] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.
- [17] I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the Bayes risk. In G. Lugosi and H. U. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory*, pages 79–93. Springer, New York, 2006.
- [18] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric quantile estimation. *J. Mach. Learn. Res.*, 7:1231–1264, 2006.